Achieving Cost/Performance Balance Ratio Using Tiered Storage Caching Techniques: A Case Study with CephFS

Michael Poat, Jerome Lauret
Brookhaven National Laboratory

Background & Problem Statement

- STAR has implemented a Ceph Distributed Storage System primarily using the POSIX compliant CephFS for processing QA, recovering DAQ files, scratch space, and backup storage.
- Can fast SSDs speed up CephFS storage?
- Goal: Balance between IO performance and cost per GB without breaking the bank.
- Ceph Cache Tiering is not a native feature of CephFS (only with Ceph object storage). M. Poat, J. Lauret – "Performance and Advanced Data Placement Techniques with Ceph's Distributed Storage System", J. Phys.: Conf. Ser. 223 – To be published.
- Can we implement a low level caching mechanism that is undetected by Ceph and give us the IO performance we desire?
- Three low-level disk caching techniques investigated (Flashcache, dm-cache, and bcache).
- Multiple disk configurations were implemented into CephFS and single and multiple thread IO performance tests were run to see the performance impact.

Analysis Procedures

- Single and multi-thread IOzone performance tests were run across all devices.
- bcache and dm-cache configurations were implemented. (Flashcache is not supported by Scientific Linux and is no longer supported natively).
- bcache, dm-cache, bare HDD & 3x HDD RAID0 CephFS clusters were benchmarked with single-thread IOzone tests and multi-threaded, multi-client IOzone tests.

STAR Ceph Distributed Storage System

Configuration

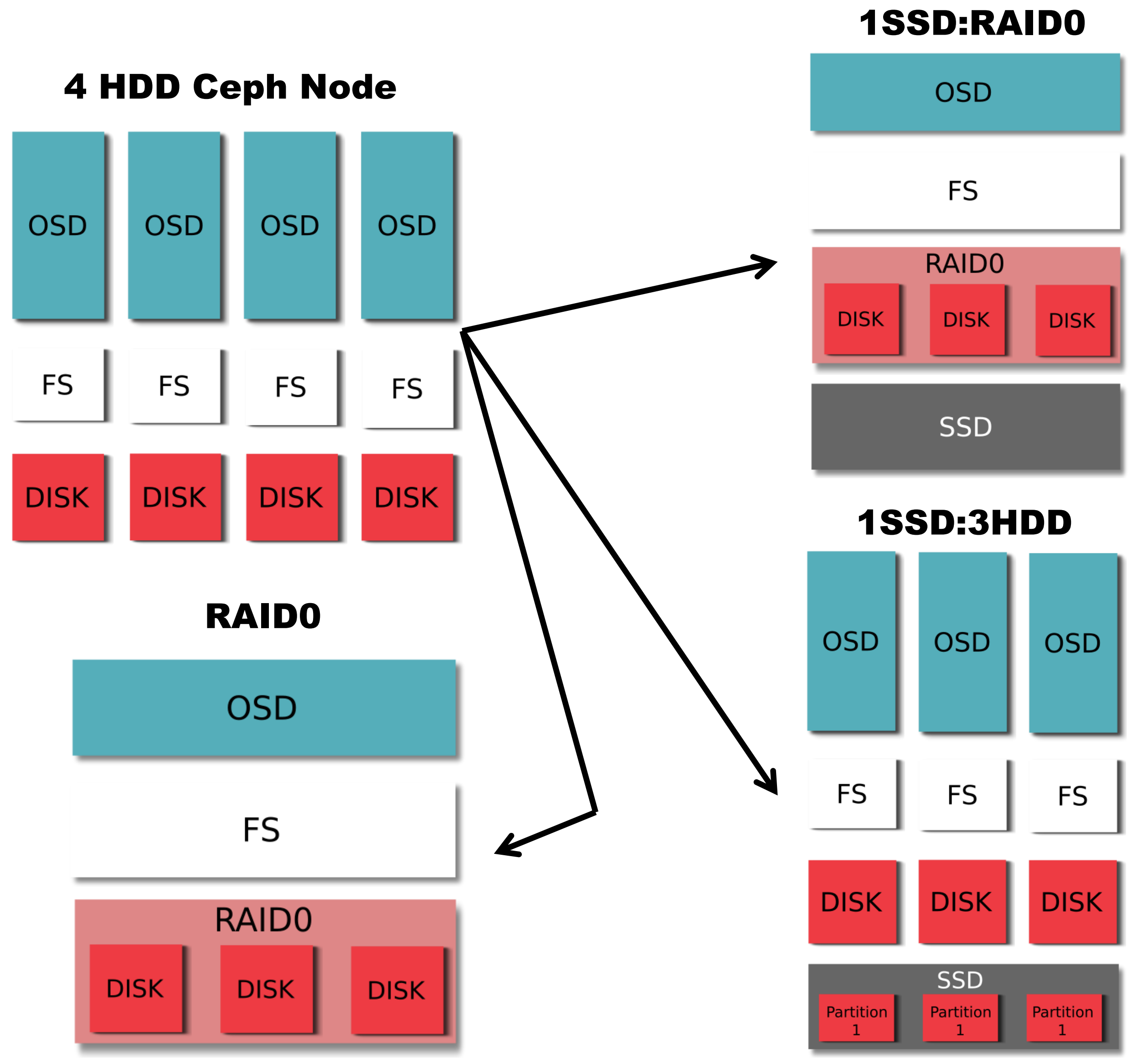
- 30 nodes with 4 – 2TB SAS HDD each.
- Replication 3, total of 80TB of redundant, fault tolerant storage.
- The primary use for Ceph is to leverage the POSIX compliant CephFS (NFS like) mountable storage for users.
- Applicable uses: processing QA, recovering DAQ files, scratch space, backup store.

Data Placement Techniques: Findings

- Ceph has built in performance based data placement techniques: OSD Pool Mapping, Primary Affinity, OSD Journals on SSDs, and Cache Tiering. Approach applied - past ACAT 2016 work.
- Efforts to replace one HDD per node with a fast drive (SSD), performance sought was not obtained.
- Performance gain is possible but at what cost?

Ceph Configurations

Stock 4 HDD Ceph Node configuration transitioned to clusters with 1SSD:3HDD, 1SSD:RAID0, and standalone RAID0 configurations.

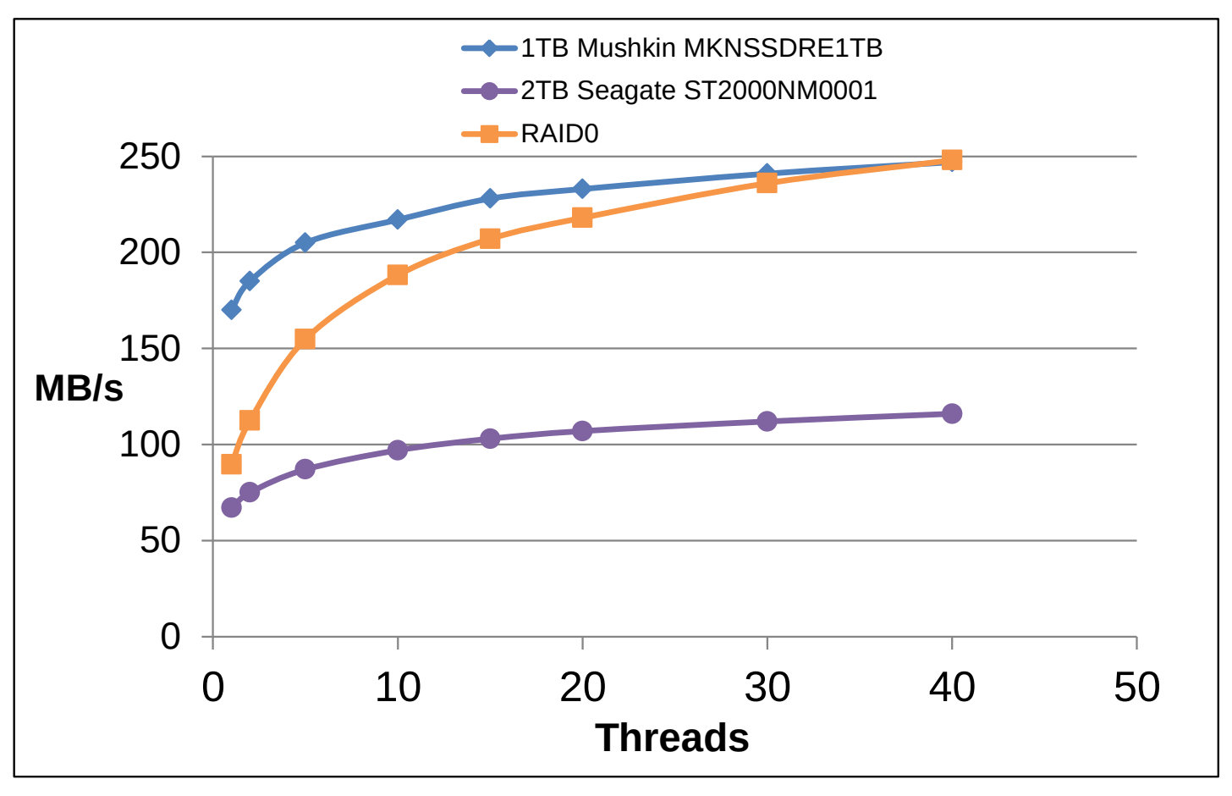


Disk Caching Techniques

Bare Disk

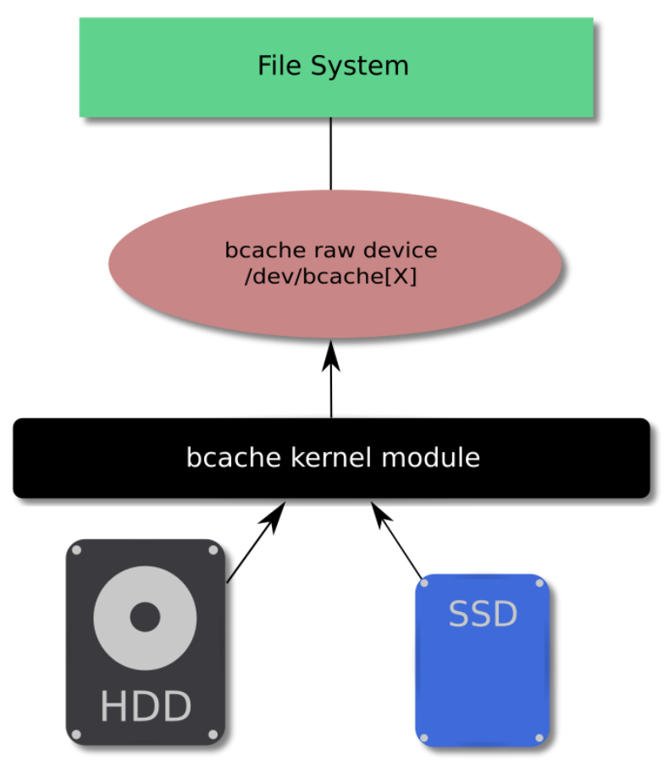
- 2TB Seagate ST2000NM0001 (HDD) and 1TB Mushkin MKNSSDRE1TB (SSD) used. RAID0 is composed of 3 HDD.
- IO Performance Test - 4096KB chunk sizes as a function of the number of threads increasing (x-axis) shown in MB/s (4096KB = Ceph block size).
- SSD performs ~2-2.5x faster than bare HDD.
- SSD outperforms RAID0 with low number of threads, near same performance at high number of threads.

IOzone – 4096KB Multi-Thread Write

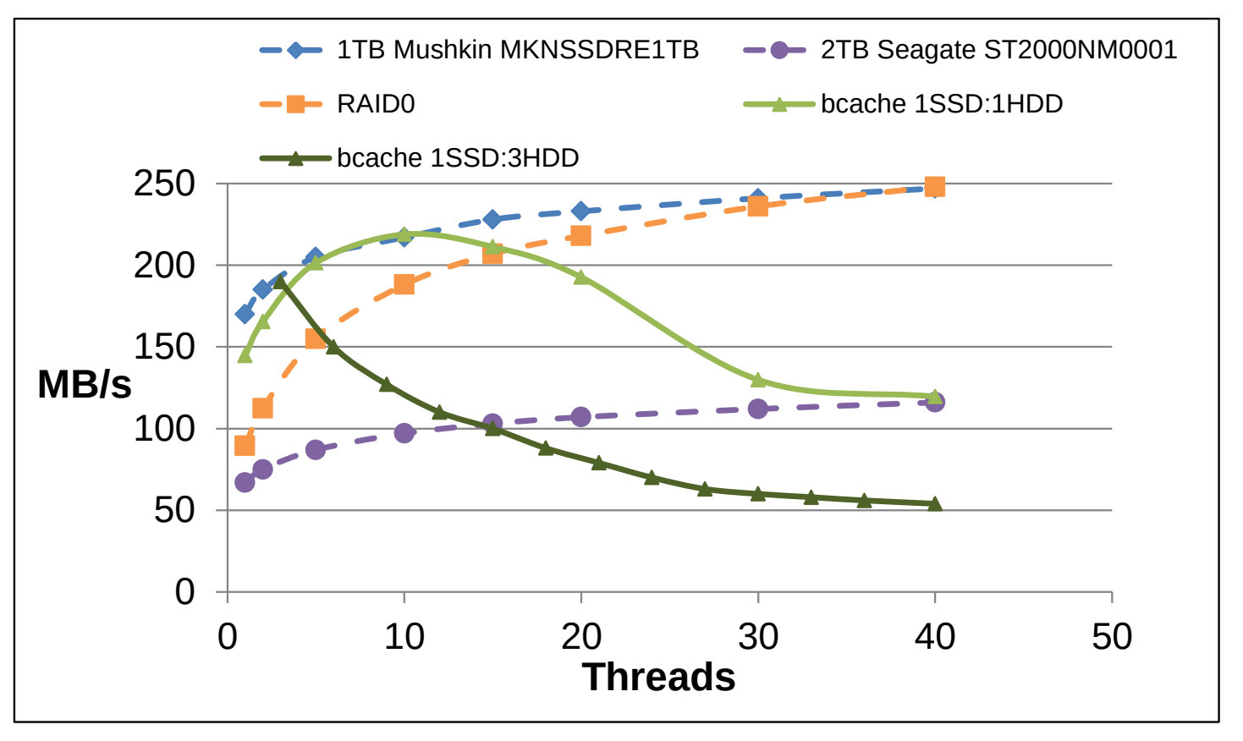


bcache

- Linux kernel block layer cache.
- One or more SSDs mapped to one or more HDDs to act as a cache.
- Green curve represents 1SSD:1HDD bcache device.
- IO performance converges with SSD at low number of threads and drops off at higher number of threads.
- 1SSD:3HDD (aggregate of all 3 devices) shows very poor performance. Not expected.
- IO is directed to SSD, bcache slows down IO under heavy load when multiple bcache devices are created.

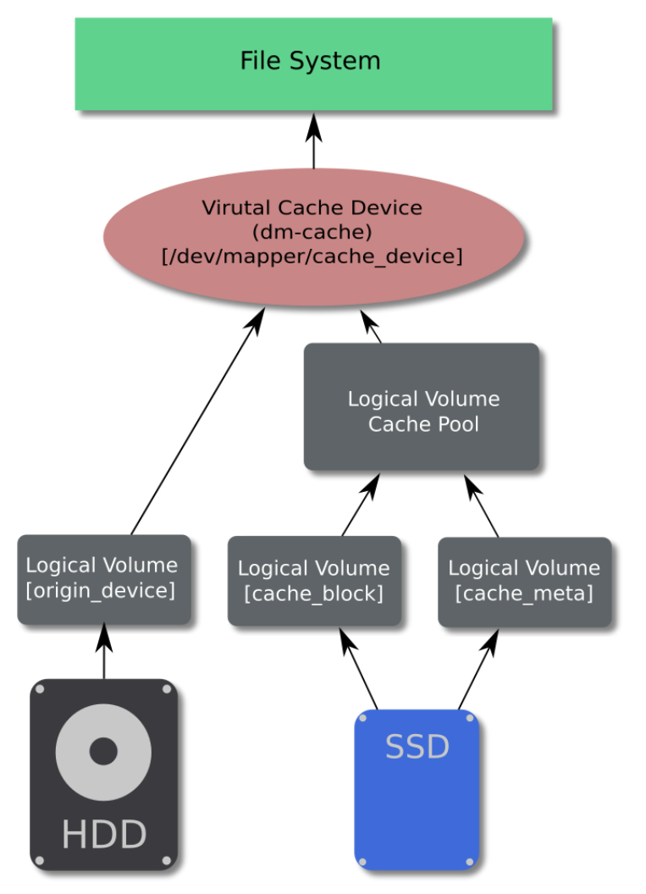


IOzone – 4096KB Multi-Thread Write

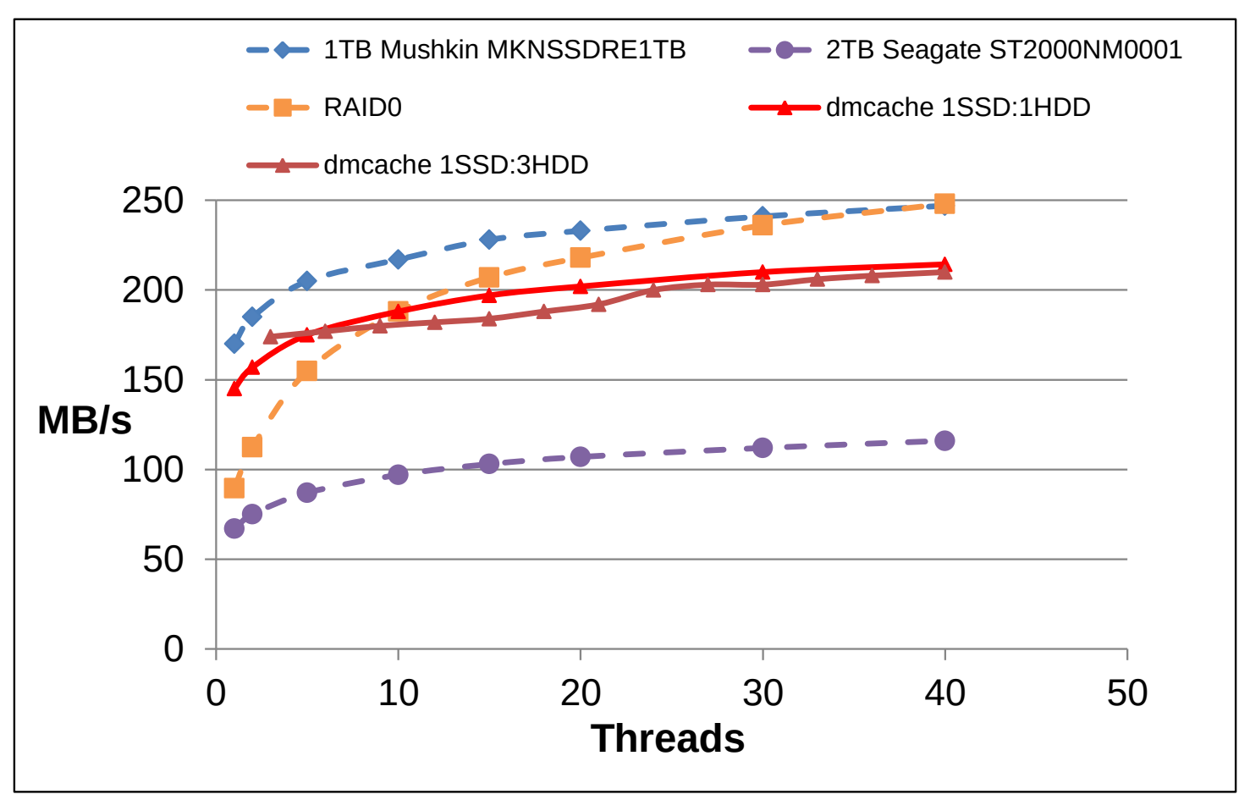


dm-cache

- Linux kernel device mapper caching technique.
- One or more SSDs can be mapped to one or more HDDs to act as a cache.
- dm-cache requires 3 logical volumes in total
  - 'Metadata' & 'Cache' Volume on SSD
  - 'Origin' Volume on HDD
- The dm-cache device is set to writeback.
- IO performance is similar with 1SSD:1HDD & 1SSD:3HDD – IO is set to write to SSD. Under heavy load, IO will writethough to backing HDDs.



IOzone – 4096KB Multi-Thread Write

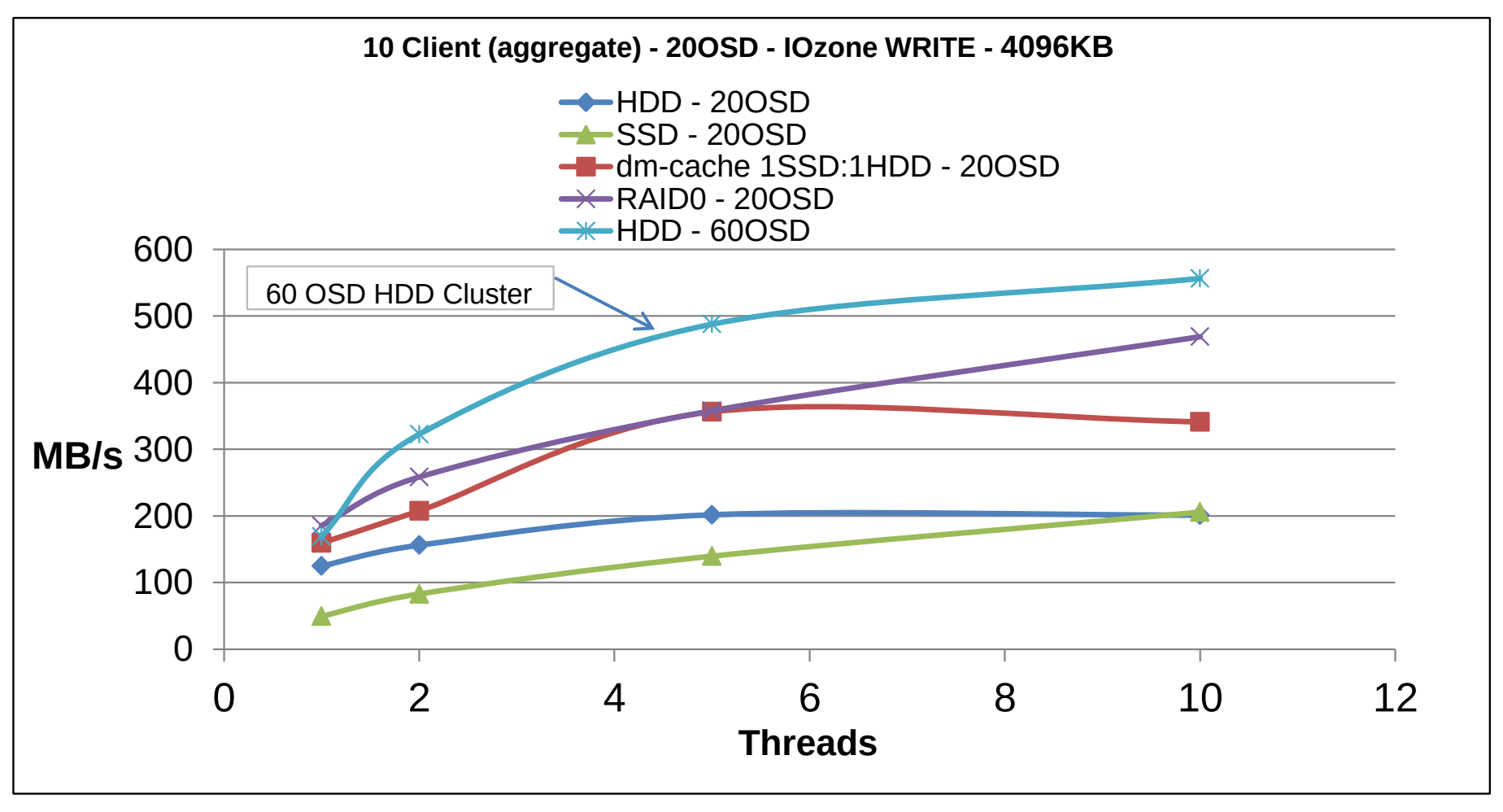


CephFS Performance Results

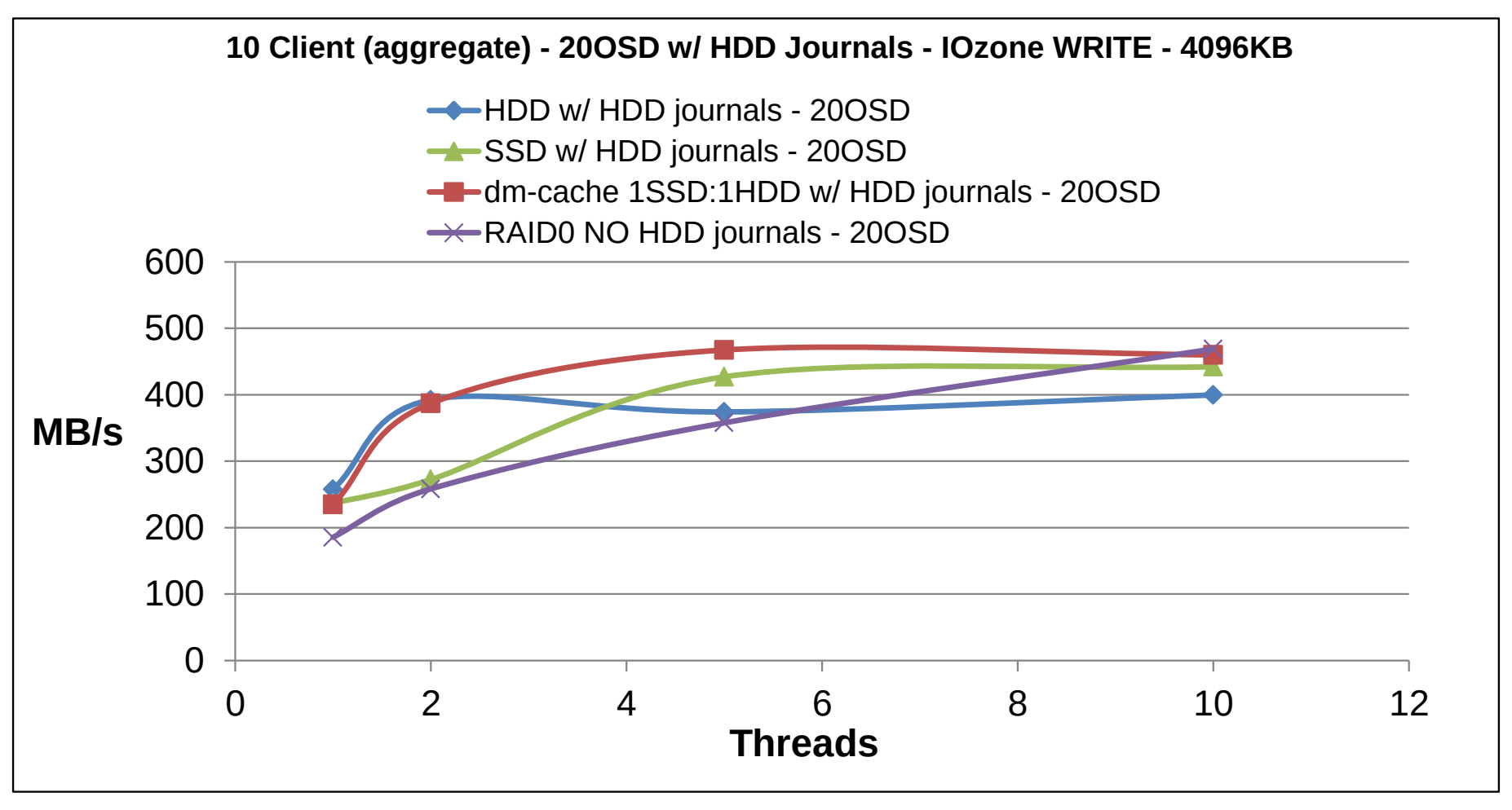
- 4 CephFS clusters made up of 20 OSDs each (HDD, SSD, 1SSD:1HDD dm-cache, & RAID0).
- 10 client IOzone 4096KB chunk writes, thread range 1-10 per client. Performance shown as aggregate.
- dm-cache performance is above SSD and HDD cluster, while the expectation would be between the two.
- dm-cache response to journal flush may be the reason for out-of-bound performance.
- SSD outperforms HDD in bare test, in Ceph context performance is flipped.
- Ceph OSD journals are bottleneck in SSD vs. HDD? →

- All Ceph OSD journals were then mounted onto a separate HDD (except RAID0)
  - ❑ SSD Ceph cluster ~3x performance increase
  - ❑ HDD Ceph cluster ~2x performance increase
  - ❑ dm-cache cluster ~0.5x performance increase
- SSD Ceph cluster with external journals = Faster IO.
- Journal write must 'sync' before proceeding to FS write. Drive dependent on ATA_CMD_FLUSH handling.
- Bare SSD Ceph cluster - lack of PLP (**P**ower **L**oss **P**rotection) cause journal flush to FS = latency.

10 Client IOzone Write into CephFS



10 Client IOzone Write into CephFS



Cost Analysis

- Current 120 2TB HDD cluster – Cost $14,400.
- In bare test - Consumer grade SSD - 4.5 times cost impact with only 2.25 times performance increase (Must test in Ceph before large purchase).
- 120 2TB Consumer SSD cluster – Cost $64,800.
- 120 2TB Enterprise SSD cluster – Cost $126,000.
- 1 Enterprise SSD:3HDD may positively impact performance but cost over HDD only cluster = $31,500 (~ x2 base cost).

Config	Avg. Speed @ 4096KB	Cost	Cost per MB/s	Cost per TB	Total space w/ 4 slot
2TB HDD	100MB/s	$120	$1.20	$60	8TB
2TB Consumer SSD	225MB/s	$540	$2.40	$270	8TB
2TB Enterprise SSD	540MB/s	$1050	$1.95	$525	8TB
dm-cache w/ ConsR. SSD + Jrnl. HDD w/ PLP	200 MB/s	$480	$2.40	$240	4TB

Conclusion

- While bare tests show performance gain using SSD over HDD, CephFS performs the best with a 'Stock' 4 HDD (120 OSD) configuration.
- RAID0 shows good performance from standalone test. In Ceph context, number of OSDs matters most. RAID0 not beneficial.
- Not all SSDs are the same, featureless SSDs may cause worse performance than HDDs in Ceph due to journaling flush-sync. Mushkin drives we used perform poorly in Ceph.
- dm-cache seems to show more stable performance than bcache. However, dm-cache would not perform well with our SSDs unless the journal is offloaded to a separate device.
- Cheap SSDs cannot help with performance gain. Enterprise models (with PLP) must be considered for performance increase. Cost for upgrade is significant.

CHEP 2016
22nd International Conference on Computing in High Energy and Nuclear Physics, Hosted by SLAC and LBNL, Fall 2016
BROOKHAVEN NATIONAL LABORATORY
U.S. DEPARTMENT OF ENERGY | Office of Science
STAR